Brügelmann, Hans

# Sind Noten nützlich - und nötig? Ziffernzensuren und ihre Alternativen im empirischen Vergleich. Eine Expertise der Arbeitsgruppe Primarstufe an der Universität Siegen im Auftrag des Grundschulverbands e. V., Frankfurt (Kurzfassung)

*Bartnitzky, Horst [Hrsg.]; Brügelmann, Hans [Hrsg.]; Hecker, Ulrich [Hrsg.]; Schönknecht, Gudrun [Hrsg.]: Pädagogische Leistungskultur. Frankfurt am Main : Grundschulverband - Arbeitskreis Grundschule e.V. 2006, S. 17-46. - (Beiträge zur Reform der Grundschule; 121)*

Hans Brügelmann mit Axel Backhaus, Sabine Bock, Erika Brinkmann (Gast),
Hendrik Coelen, Thomas Franzkowiak, Simone Knorre,
Barbara Müller-Naendrup, Elisabeth Oser, Sara Roth

# *Sind Noten nützlich – und nötig?*

## *Ziffernzensuren und alternative Formen der Leistungsbeurteilung im empirischen Vergleich*

*Eine Expertise*[1] *der Arbeitsgruppe Primarstufe an der Universität Siegen im Auftrag des Grundschulverbands e.V., Frankfurt*

*– Kurzfassung –*

## *0. Auftrag und Kontext der Expertise*

Noten werden unter den Betroffenen, aber auch in Fachkreisen immer wieder kontrovers diskutiert. Ihr Nutzen war und ist heftig umstritten. Angesichts der ungebrochen harten Auseinandersetzungen ein erstes frappierendes Ergebnis unserer Literaturrecherche: Zentrale empirische Befunde zur Problematik von Noten liegen seit 50 Jahren, zum Teil noch länger vor. Seit den 1970er Jahren sind diese Studien im deutschen Sprachraum in systematisierenden Überblicken publiziert worden. Ihre Kritik ist in der Zwischenzeit durch weitere Studien bestätigt und erneut mehrfach zusammengefasst worden. Dennoch hat sich in der Wahrnehmung von SchülerInnen, Eltern und LehrerInnen, aber auch durch die Öffentlichkeit und erst recht im Schulalltag nur wenig verändert. Soweit es in den letzten Jahren Veränderungen gab, sind – vor allem im Grundschulbereich – eher Tendenzen zu beobachten, das Rad der Entwicklung zurückzudrehen, ja, die Noten über den Leistungsbereich hinaus auszuweiten.

Vor diesem Hintergrund hat der Grundschulverband die vorliegende Expertise in Auftrag gegeben. Sie soll die empirische Forschung zu Ziffernnoten und alternativen Formen der Leistungsbeurteilung sichten und im Kontext der aktuellen Diskussion bewerten. Im Fokus des Gutachtens steht die Grundschule. Viele der ausgewerteten Studien und auch viele unserer Überlegungen beziehen sich aber auf grundsätzliche Fragen der Leistungsbeurteilung und reichen deshalb über diese Schulstufe hinaus.

---

1 Der folgende Text ist die Kurzfassung des ausführlichen Gutachtens, in dem die einschlägigen Publikationen, insbesondere die empirischen Studien, differenziert ausgewertet und belegt sind. Verweise im Text beziehen sich auf die entsprechenden Kapitel der Langfassung.
Unsere Analysen legen grundlegende Probleme einer pädagogischen Leistungsbeurteilung offen, die Folgerungen beziehen sich zunächst einmal aber vor allem auf die Grundschule.

## 0.1 Ansatz und Aufbau des Gutachtens

Ziffernnoten werfen sehr unterschiedliche Probleme auf. Drei Entscheidungsebenen mit je besonderen Problemen sind zu unterscheiden:

- die Wahl der Verfahren zur *Feststellung* des Lernerfolgs: informelle vs. standardisierte Aufgaben, offene vs. strukturierte Beobachtung;
- die Wahl der Bezugsnorm zur *Bewertung* des Lernerfolgs: nach Annäherung an das Lernziel und / oder individuellem Lernforschritt und / oder relativer Leistungsposition in einer Gruppe;
- die Wahl der *Darstellungsform* in der Rückmeldung: Beschreibung vs. Bewertung; freie Formulierung vs. Ziffern.

Die genannten Alternativen werden im Schul- und Berufsalltag in verschiedenen Kombinationen realisiert. Dadurch wird eine sachliche Diskussion erschwert. Sorgfältig zu trennen sind nämlich mehrere Teilfragen:

*In welcher Funktion werden Leistungen beschrieben und bewertet?* (→ Kap. 0.3 und 7)
Leistungen können im Blick auf einen festzustellenden Förderbedarf beurteilt werden oder auch, um den Unterricht zu verbessern. Im deutschen Schulsystem dominieren dagegen die Selektions- und Disziplinierungsfunktion. Dieser institutionelle Kontext prägt die Wirkung von Noten – und schränkt die Möglichkeiten alternativer Beurteilungsformen ein.

*Über welche Verfahren werden Leistungen erfasst?* (→ Kap. 1)
Noten wird vorgeworfen, sie seien nicht objektiv, nicht valide und nicht zuverlässig. Diese Probleme haben aber auch Verbalzeugnisse. Beider Datengrundlage ist an die Person der Beurteilenden und ihre Auswahl der Instrumente zur Erhebung von Leistungen gebunden. Insofern sind als Alternative zu Klassenarbeiten und informellen Beobachtungen standardisierte Tests und strukturierte Beobachtungen zu diskutieren.

*Anhand welcher Maßstäbe werden Leistungen bewertet?* (→ Kap. 2)
Noten wird eine einseitige Orientierung an der sozialen Bezugsnorm – mit der jeweiligen Schulklasse als dominierendem Maßstab – vorgeworfen. Diese Verbindung ist aber nicht zwingend. Noten können sich auch am Lernfortschritt oder an den Anforderungen orientieren (und sollen dies sogar, vgl. bereits KMK 1968). Umgekehrt orientieren sich auch Verbalzeugnisse nicht zwangsläufig an der individuellen Entwicklung. Die Bedeutung und die Wirkungen unterschiedlicher Maßstäbe für die Bewertung von Leistungen sind also übergreifend zu klären.

*In welchen Formen werden Leistungsbeurteilungen dargestellt?* (→ Kap. 3)
Erst auf dieser Stufe geht es um Ziffernnoten vs. sprachliche Formulierungen. Dabei interessieren vom Gutachtenauftrag her zwei Fragen:

- Werden die Ansprüche der beiden Zeugnisformen in der praktischen Umsetzung tatsächlich eingelöst? (→ 3.1)

➢ Welche Wirkungen haben verschiedene Rückmeldeformate auf den Unterricht bzw. auf die Entwicklung der SchülerInnen (Erfüllung verschiedener Erwartungen / Funktionen und etwaige negative Nebenwirkungen)? (→ 3.2)

*Wie werden verschiedene Zeugnisformen wahrgenommen?* (→ Kap. 4)
Wer Beurteilungsformen ändern will, muss deren Akzeptanz und etwaige politische Vorbehalte bzw. persönliche Bedenken und Ängste kennen. Unabhängig von den empirisch festgestellten Stärken und Schwächen verschiedener Formate geht es darum, wie die Beteiligten selbst die Leistungsfähigkeit unterschiedlicher Darstellungsformen einschätzen. Im Vordergrund steht die Frage, wie gut die Formate die unterstellten Funktionen – nach Einschätzung verschiedener Gruppen – erfüllen (insbesondere Informationsgehalt und Verständlichkeit der Rückmeldung).

*In welchem Verhältnis stehen Aufwand und Ertrag verschiedener Darstellungsformen?* (→ Kap. 5)
Die Chancen von Reformen hängen schließlich auch davon ab, dass sie von den Beteiligten nicht nur als inhaltlich wichtig, sondern auch als praktikabel, zumindest nicht als unergiebige zusätzliche Belastung, wahrgenommen werden.

## *0.2 Datengrundlage des Gutachtens*

Die Fragestellung des Gutachtens ist so komplex und das Datenmaterial so heterogen, dass eine rein statistische Metaanalyse von Daten aus verschiedenen Studien nicht in Frage kommt. Befunde liegen vor aus Laborversuchen, aus Feldexperimenten sowie von Beobachtungen und Befragungen unter nicht kontrollierten Bedingungen. Diese unterschiedlichen Datenformate verlangen eine qualitative Analyse und Interpretation. Der Gefahr einer zu stark personabhängigen Auswahl und Bewertung der Studien haben wir durch zwei Schritte der sozialen Kontrolle entgegengewirkt: Schon das Gutachten selbst wurde nicht von einer Person, sondern im Team erstellt und im mehrfachen Austausch wechselseitig kommentiert. Zusätzlich zu dieser gruppeninternen Diskussion wurde das Gutachten externen ExpertInnen[2] zur Kommentierung vorgelegt.

Erfreulicherweise können wir auf eine Reihe aktueller Studien im deutschsprachigen Raum zurückgreifen, bei denen die Notenfrage (meist im Vergleich mit Verbalgutachten) im Zentrum der Untersuchung steht oder einzelne ihrer Aspekte im Zusammenhang mit anderen Fragen (z. B. Lehrerurteil vs. Testwerte) behandelt werden. Anhand von Daten aus den internationalen Studien TIMSS, PISA und IGLU können die Organisation und Form von Prüfungen zudem im Kontext und Vergleich unterschiedlicher Bildungssysteme analysiert werden.

---

2 Wir danken Heide Bambach, Horst Bartnitzky, Wolfgang Harder, Peter Heyer, Hans Werner Heymann, Georg Lind, Heide Niemann, Argyro Panagiotopoulou, Markus Roos, Corinna Schmude, Christop Selter und Felix Winter sowie den Mitgliedern des OASE-Kolloquiums an der Universität Siegen für kritische Anmerkungen und hilfreiche Anregungen zu dieser Kurzfassung in unterschiedlichen Stadien der Arbeit.

Unsere Literatursuche in internationalen Datenbanken zu Stichworten wie ›assessment‹, ›marking‹ und ›grading‹ war dagegen wenig ergiebig. Auf den ersten Blick überrascht dies, werden doch in den angelsächsischen Ländern schon länger als in Deutschland alle möglichen Aspekte von Unterricht immer wieder empirisch untersucht. Zu bedenken ist aber, dass die Notendiskussion nicht (mehr) überall den gleichen Stellenwert hat. In vielen westlichen Industrieländern besteht eine langjährige Tradition, für die Leistungsbeurteilung standardisierte Tests einzusetzen. Zum Teil ist dies eine Folge der empirischen Kritik an Noten vor und nach dem zweiten Weltkrieg. Zum Teil richtet sich diese Kritik – in anderer Form – heute gegen standardisierte Tests selbst. So wird mehrfach gefordert, das Gewicht des Lehrer-Urteils wieder zu stärken. Hilfreich für unser Gutachten waren insofern neuere *reviews* zur Wirkung verschiedener Formen der Leistungsbewertung in England und den USA. Ihre Ergebnisse und Folgerungen decken sich in erstaunlich hohem Maße mit unserem Resümee der deutschsprachigen Studien, deren Ergebnisse unter ganz anderen Rahmenbedingungen gewonnen wurden.

## *0.3 Historischer Rückblick und gesellschaftlicher Kontext*

Verbale Beurteilungen gab es, wenn auch nicht in der heutigen Form, wesentlich früher als Ziffernzeugnisse. Deren Einführung war mit dem Anspruch verbunden, das Leistungsprinzip gegen die Verteilung gesellschaftlicher Positionen nach sozialer Herkunft durchzusetzen. Neben emanzipativen Wirkungen erfüllten sie aber auch immer eine Ausgrenzungsfunktion. Nur geprüfte, d. h. institutionell zertifizierte Leistungen zählen.

An Noten werden sehr unterschiedliche Erwartungen gestellt:

- Förderung des Lernens durch Motivations-, Disziplinierungs- und Sozialisationseffekte,
- Rückmeldung an die SchülerInnen und ihre Eltern über den Erfolg individueller Lernprozesse und des Unterrichts insgesamt,
- Ausweis von Leistungen für Selektionsentscheidungen.

Diese Vielfalt kann in der Praxis zu einer Funktions-Überlast und im Ergebnis zu einer einseitigen Betonung einzelner Funktionen führen. Aber auch die Diskussion über den Sinn von Noten wird dadurch belastet, dass sich verschiedene Fragen mischen, z. B. nach der Funktion und dem Zeitpunkt von Prüfungen generell, nach ihrer Datengrundlage und Darstellungsform andererseits (→ Kap. 0.1).

Ziffernnoten sind bereits in der Reformpädagogik Gegenstand heftiger Kritik gewesen. Als Alternative wurden in den 1970er Jahren auch von der Bildungspolitik ausdifferenziertere Diagnosebögen und Verbalgutachten – zumindest für die ersten Grundschuljahre – favorisiert. Dieser Trend hat sich in den letzten Jahren teilweise wieder umgekehrt. In der Didaktik dagegen geht die Entwicklung weit über diese Kontroverse hinaus. Gefordert werden gehaltvollere Dokumentationsformen wie Lerntagebücher oder Portfolios und vor allem dialogische Formen, in denen Fremdbewertungen durch Selbsteinschätzungen ergänzt werden, die dann auch in konkrete Zielabsprachen für die Arbeit im Unterricht münden.

## *0.4 Die Situation in den Bundesländern: ein Überblick*

In den 1970er und 1980er Jahren ist der notenfreie Raum in den Grundschulen schrittweise erweitert worden. Aber inzwischen dominieren Noten – nach rein verbalen Beurteilungen im Anfangsunterricht – wieder ab Klasse 3, zum Teil schon Ende der zweiten Klasse. Seit Mitte der 1990er Jahre wurden die generelle Notenfreiheit in den ersten Klassen wie auch die Möglichkeiten für Ausnahmen von der Ziffernbewertung in den höheren Klassen schrittweise wieder zurückgenommen. Zugleich wird der Anwendungsbereich der Noten wieder erweitert (Frühenglisch; ›Kopfnoten‹ für das Sozial- und Arbeitsverhalten).

Gegenläufig zu diesen Entwicklungen finden alternative Formen der Leistungsbewertung auf der Sekundarstufe und im tertiären Bereich wachsende Aufmerksamkeit – bis hin zur Verbreitung notenfreier Beurteilungsverfahren in der Berufswelt. Insofern ist das Argument, Ziffernnoten seien notwendig, um SchülerInnen auf den ›Ernst des Lebens‹ vorzubereiten, überholt.

## *0.5 Blicke über den Zaun: die internationale Situation*

Schon in Europa ist die Situation sehr heterogen. Im Vergleich zu Deutschland sind allerdings in zahlreichen Ländern (bis auf den ehemaligen Ostblock) Selektionsentscheidungen wie Zurückstellung, Sitzenbleiben, Überweisung in Sonderschulen wesentlich seltener, dauert der gemeinsame Unterricht länger und setzt auch eine Benotung von Leistungen später ein. Viele dieser Länder schneiden bei internationalen Vergleichen besser ab als Deutschland.

Übersichtliche Muster und einfache Abhängigkeiten gibt es aber nicht. Dazu sind die Konstellationen zu komplex und vielfältig (z. B. Beginn der Schulpflicht, Dauer der gemeinsamen Schulzeit, System/e der Leistungsbeurteilung, Formen und Umfang der Selektion einerseits sowie Abschneiden in verschiedenen Fächern bei IGLU bzw. PISA andererseits). Festhalten lässt sich aber:

- Es gibt mehrere Länder mit späterer Benotung und weniger Selektion, die in den Leistungsvergleichen deutlich besser abschneiden als Deutschland (vgl. etwa Schweden oder Südtirol).
- In dieser Gruppe finden sich aber auch Länder, die nicht besser oder sogar schlechter abschneiden als Deutschland (z. B. Norwegen).
- Schließlich gibt es innerhalb Deutschlands auch Länder mit früherer Benotung und / oder stärkerer Selektion (z. B. Bayern und Baden-Württemberg), die zumindest oberflächlich erfolgreicher zu sein scheinen als Bundesländer, die bis vor kurzem noch stärker integrative Ansätze favorisiert haben (z. B. Nordrhein-Westfalen oder Bremen).

Fazit: Strukturelle Systemmerkmale garantieren keine pädagogischen Erfolge. Die Abschaffung von Noten ist kein ›Selbstläufer‹. Nicht die Verwendung von Zif-

fern, sondern das pädagogische Ethos im Unterricht und insbesondere das Gewicht von Klassifikation und Selektion machen die entscheidende Differenz aus (→ Kap. 7). Dieses Ergebnis deckt sich mit dem später zu erörternden Befund, dass auch deutschlandintern die Wirkungen von Noten vs. Verbalgutachten innerhalb der beiden Ansätze stärker streuen als zwischen ihnen. Allerdings machen die internationalen Vergleiche deutlich: Ein Verzicht auf Benotung und Selektion in den ersten Schuljahren und über die Grundschulzeit hinaus ist kein Hindernis für eine erfolgreiche pädagogische Arbeit – auch im fachlichen Leistungsbereich.

## *1. Mit welchen Verfahren werden Leistungen erfasst?*

Die Fundiertheit von Beurteilungen hängt von ihrer Datengrundlage ab. Forschungsmethodisch wird deren Qualität in der Tradition quantitativ-standardisierter Forschung über drei Gütekriterien bestimmt:

- Gültigkeit (›Validität‹),
- Personunabhängigkeit (›Objektivität‹) und
- Verlässlichkeit (›Reliabilität‹).

Zu wenig Beachtung finden bisher Kriterien aus der allgemeinen Evaluationsdiskussion wie Fairness, Glaubwürdigkeit, Stimmigkeit, Ökonomie, Nützlichkeit.

### *1.1 Wie gut erfassen Leistungsbeurteilungen, was sie erfassen sollen? (Validität)*

Leistungen sind beobachtbare Verhaltensweisen. Ihre Beurteilung zielt aber nicht nur auf das beobachtete Verhalten (›Performanz‹), sondern auch auf die zugrunde liegenden Fähigkeiten (›Kompetenz‹). Die Gültigkeit einer Leistungsbeurteilung ist nur schwer zu begründen, da die Außenkriterien selbst umstritten sind.

#### *1.1.1 Wie gut sind die Kriterien für Leistungsbeurteilungen inhaltlich abgesichert?*

Die Diskussion über die Bildungsstandards zeigt ein hohes Maß an Uneinigkeit, was als Mindest- oder Regelleistung eingefordert werden kann. Die Kritik an den Aufgaben der landesweiten Lernstandserhebungen und internationalen Leistungsvergleiche hat offengelegt, wie umstritten die Annahmen zu den angeblich erfassten ›Fähigkeiten‹ sind. So führen Tests aus demselben Bereich, z. B. in Mathematik oder Rechtschreibung am Ende der Grundschulzeit, zu unterschiedlichen Leistungseinschätzungen, weil sie inhaltlich unterschiedliche Schwerpunkte setzen. Dieselbe Fähigkeit (z. B. ›Lesekompetenz‹) wird über unterschiedliche Aufgaben operationalisiert. Entsprechend unterschiedlich fällt die Einschätzung der Lesefähigkeit deutscher SchülerInnen etwa bei PISA und DESI aus.

Die Validität von Noten wird in den letzten Jahren oft mit dem Hinweis kritisiert, dass Noten nicht gut mit den Ergebnissen von Leistungstests in den entsprechenden Fächern übereinstimmen. Andererseits wird die inhaltliche Gültigkeit

von Tests in der Regel auch damit begründet, dass ihre Ergebnisse in der Normierungsphase ›gut‹ mit den Lehrerurteilen übereinstimmen. Damit entstehen Kreisschlüsse, bei denen kein Verfahren beanspruchen kann, besser zu sein als das andere. Das einzige unabhängige Kriterium ist ihre Vorhersagekraft, bezogen auf zukünftige Leistungen. Diese aber erweist sich als sehr begrenzt (s. dazu unten → 1.1.3).

*1.1.2 Wie gut stimmen Beurteilungen aus verschiedenen Quellen überein?*

Nicht nur die Ergebnisse von verschiedenen Tests desselben Bereichs, auch Fachnoten und Tests stimmen nur begrenzt überein. Problematisch an der Diskussion ›nach PISA‹ ist, dass Tests dabei fast selbstverständlich als Maßstab für die ›wahre‹ Leistung von SchülerInnen gesetzt werden. Vergleiche mit weiteren Kriterien zeigen aber, dass Lehrerurteil und Tests unterschiedliche Aspekte fachlicher Leistungen erfassen. Es macht deshalb keinen Sinn, die Qualität des einen Verfahrens allein durch den Grad der Übereinstimmung mit den Ergebnissen des anderen zu bestimmen.

Aber selbst wenn man die Testergebnisse als Maßstab nimmt, sind für die beobachteten Abweichungen immer noch zwei verschiedene Erklärungen denkbar:

➢ LehrerInnen erkennen schlechter als Tests, wo Kinder in ihrer Entwicklung stehen, welche Schwierigkeiten sie bei der Auseinandersetzung mit dem jeweiligen Gegenstand haben (fehlende diagnostische Kompetenz).

Oder:

➢ LehrerInnen verschiedener Klassen ordnen den richtig erkannten Leistungsstand unterschiedlichen Notenstufen zu (abweichende Bewertungsmaßstäbe).

Die vorliegenden Studien sprechen eher für die zweite Sicht. So korrelieren Noten und Testwerte *innerhalb* von Klassen sehr viel höher miteinander als über verschiedene Klassen hinweg. LehrerInnen differenzieren unterschiedliche Lernstände also auch im Sinne von Leistungstests einigermaßen zutreffend, aber sie setzen anschließend den Bezugspunkt für die Bewertung durch Noten unterschiedlich an. Bestätigt wird diese Deutung durch noch höhere Korrelationen, wenn man von den LehrerInnen *qualitative* Urteile über die voraussichtliche Entwicklung ihrer SchülerInnen erfragt und diese mit Tests zur kognitiven Leistungsfähigkeit abgleicht.

*1.1.3 Wie genau lässt sich aus der Beurteilung von Leistungen deren zukünftige Entwicklung vorhersagen (prognostische Validität)?*

Leistungen – als beobachtbare Verhaltensweisen – werden rückblickend beurteilt. Die Beurteilung soll aber die zugrunde liegende Fähigkeit erfassen und damit auch Aufschluss geben über zukünftig zu erwartende Leistungen. Die Vorhersagekraft von Beurteilungen, insbesondere von Noten, ist in verschiedenen Phasen der Bildungslaufbahn untersucht worden.

*1.1.3.1 Kindergarten → Schulerfolg*

Schon die Prognosen des Schulerfolgs im Kindergarten haben eine hohe Fehlerquote. Diese Einsicht hat zur Abschaffung der Schulfähigkeitstests geführt, die zunächst durch Beobachtungsverfahren ersetzt wurden. In einem weitergehenden Schritt wurde in einigen Bundesländern die Selektionsentscheidung selbst abgeschafft, da sich gezeigt hat, dass sich Kinder mit demselben Testergebnis / Reife- bzw. Fähigkeitsurteil je nach den Bedingungen im Anfangsunterricht ganz unterschiedlich entwickeln können. Auch Klassifikationsversuche mit Hilfe fachbezogener Verfahren haben eine zu hohe Fehlerquote. Im Bereich der Schriftsprache schwankt sie zum Beispiel für die fonologische Bewusstheit – je nach Verfahren, Zeitspanne der Prognose und vor allem Art des zwischenzeitlichen Unterrichts – zwischen 20% und 80%. Bei derart hohen Fehlprognosen lassen sich keine Fördermaßnahmen, erst recht aber keine Selektionsentscheidungen rechtfertigen – ein Befund, der auch beim Einsatz von Sprachstandserhebungen vor der Schule zu beachten ist.

*1.1.3.2 Schule → Fachleistungen über Schuljahre hinweg*

Die Fehleranfälligkeit von Vorhersagen zeigt sich auch bei Prognosen von einem Schuljahr zum nächsten. Zwar liegen Korrelationen für die Übereinstimmung von Tests innerhalb eines Faches bei .60 bis .70[3]. Diese als hoch erachteten Zusammenhänge in einer Gruppe verdecken aber zum Teil erhebliche Positionsveränderungen auf der Einzelfallebene. Bei Vorhersagen über mehrere Schuljahre, also mit zunehmender Dauer, sinken die Korrelationen zudem.

Empfehlungen von LehrerInnen für die Wahl weiterführender Schulen haben ein entsprechend hohes Fehlerrisiko. Es fehlen überdies Belege, dass Empfehlungen eine größere Prognosesicherheit beanspruchen können als die freie Wahl der Eltern. Sichtbar wird die generelle Unsicherheit von Vorhersagen darin, dass sich die Leistungsverteilungen von Gymnasium, Haupt- und Realschule erheblich überlappen. Sowohl von Tests als auch von den Empfehlungen der LehrerInnen sind keine verlässlicheren Vorhersagen des Schulerfolgs zu erwarten, als wenn Eltern nach fachlicher Beratung über die Wahl der Schule entscheiden.

*1.1.3.3 Schule → Studien-/Ausbildungserfolg*

Noch lockerer ist der Zusammenhang zwischen den Noten bzw. Tests zum Abschluss der Schulzeit und dem späteren Studien- bzw. Ausbildungserfolg der AbsolventInnen. Die Korrelationen liegen mit .20 bis .50 deutlich niedriger als die schulinternen Zusammenhänge – sowohl bei unterrichtsbegleitenden Beobachtungen und informellen Prüfungen (wie in Deutschland) als auch beim Einsatz standardisierter Tests (wie z. B. in den angelsächsischen Ländern).

*1.1.3.4 Studium / Ausbildung→ Berufserfolg*

Kaum mehr aussagekräftig sind die Beziehungen zwischen den Abschlussnoten der Ausbildung und dem Berufserfolg. Die Korrelationen liegen zunächst bei .10

3 Eine Korrelation von 1.0 drückt aus, dass die Werte oder zumindest die Rangplätze in einem Test vollständig durch die Werte bzw. Rangplätze im anderen vorhergesagt werden können. Dagegen streuen die Werte beider Tests bei einer Korrelation von .0 völlig unabhängig voneinander.

bis .30, sowohl was das Risiko von Arbeitslosigkeit betrifft als auch bezogen auf die Berufsposition, das Einkommen und die persönliche Zufriedenheit der Erwerbstätigen. Schon nach wenigen Jahren tendieren die Zusammenhänge sogar gegen .0 – stärker bei Noten auf der Basis von Klausuren und standardisierten Tests als bei Prüfungsformen, die komplexere Leistungen anfordern (wie mündliche Prüfungen). Bei der Auswahl von BewerberInnen für berufliche Aufgaben haben sich qualitative lernbiografische Daten meist als aussagekräftiger erwiesen als punktuelle Prüfungen.

#### *1.1.4 Zwischenbilanz zu ›Validität‹*

Sowohl das Lehrerurteil als auch standardisierte Tests haben nur eine beschränkte Aussagekraft, erst recht, wenn aus beobachteten Leistungen auf Fähigkeiten einer Person geschlossen werden soll. Vor allem für Selektionsentscheidungen sind sie nicht valide genug, um den weiteren Lernerfolg einzelner Personen zureichend sicher vorhersagen zu können. Die Entwicklung von Personen ist nicht berechenbar – und variiert vor allem in Wechselwirkung mit den Lernbedingungen.

## *1.2 Wie unabhängig sind Beurteilungen von persönlichen Einflüssen? (Objektivität)*

Wenn lern- und berufsbiografische Entscheidungen durch Leistung und nicht durch Herkunft oder persönliche Beziehungen bestimmt werden sollen, muss die Leistungsbewertung unabhängig von der beurteilenden Person sein.

#### *1.2.1 Objektivität des Lehrerurteils*

Noten und andere Formen der Einschätzung von Leistungen sind in hohem Maße personabhängig. Eine bewusste Empathie hat zwar Vorteile für förderorientierte Rückmeldungen. Subjektivität ist insofern die Basis einer ermutigenden Rückmeldung. Denn diese setzt die Bereitschaft und Fähigkeit voraus, sich in die Probleme einer Person, die weniger Kompetenz als der Beurteilende hat, einzufühlen, und ist insofern Ausdruck pädagogischen Taktes im Umgang mit ihrer besonderen Verletzlichkeit.

Fatal wirken sich dagegen unterschiedliche Maßstäbe und persönliche Sympathie oder der Einfluss von sachfremden Informationen bei Selektionsentscheidungen aus. Wie viele Untersuchungen zeigen, können Noten für dieselbe Arbeit über vier bis sechs (!) Stufen streuen. Zusätzlich sind systematische Verzerrungen vor allem für die Merkmale ›Geschlecht‹, ›soziale Herkunft‹, ›Migrationshintergrund‹ und ›Verhaltensauffälligkeiten‹ nachgewiesen.

#### *1.2.2 Kann der Einsatz standardisierter Tests das Objektivitätsproblem lösen?*

Als Ausweg wird eine Standardisierung der Aufgaben, ihrer Durchführung und Auswertung propagiert. Oberflächlich erreicht man dadurch eine Eindeutigkeit der Erhebungsdaten – allerdings auf Kosten eines neuen Problems: Verhalten ist mehrdeutig und deshalb interpretationsbedürftig. Dieses Problem stellt sich bei allen Formen der Leistungsbeurteilung, macht sich aber verschärft bei stan-

dardisierten Tests bemerkbar. Wie Prüflinge eine Aufgabe gedeutet und wie sie ihre Antworten gemeint haben, ist durch die bewusste Ausblendung persönlicher Interaktionen nicht mehr verhandelbar. Damit wird nicht Objektivität gesichert, sondern die Subjektivität der TestentwicklerInnen und -auswerterInnen über die der beurteilten Personen privilegiert.

#### *1.2.3 Wie weit lässt sich das Lehrerurteil objektivieren?*

Zwei Formen der Objektivierung sind denkbar: methodisch-technisch durch die inhaltliche Präzisierung von Kriterien und Maßstäben bzw. sozial durch den kollegialen Austausch und die wechselseitige Kontrolle mehrerer PrüferInnen. Beide Maßnahmen können die Streubreite der Urteile reduzieren. Abweichungen bleiben aber auch dann in einem immer noch bedeutsamen Maße bestehen. Angelsächsische Studien verweisen auf die Notwendigkeit, im Bewertungsverfahren drei Elemente zu stärken:

- klare Definition von Kriterien,
- deren gemeinsame Erarbeitung mit den AnwenderInnen und ihre
- kontinuierliche Verfeinerung durch einen wechselseitigen Austausch in der Anwendung.

#### *1.2.4 Zwischenbilanz zu ›Objektivität‹*

Unterschiedliche Maßstäbe, aber auch sachfremde Gesichtspunkte wie Sprachstil oder Sozialverhalten des Schülers bzw. persönliche Sympathien der Lehrperson beeinflussen das fachbezogene Urteil und schränken deshalb die Objektivität sowohl von Noten als auch von Verbalgutachten erheblich ein. So wichtig das Bemühen ist, Willkürlichkeit in der Bewertung auszuschließen – die Bedeutung von Empathie für eine lernförderliche Leistungsbeurteilung darf darüber nicht vergessen werden. Nachgewiesen sind aber auch systematische Verzerrungen durch Gruppenmerkmale wie Geschlecht, soziale Herkunft und ethnische Zugehörigkeit. In Tests werden deshalb Aufgaben, ihre Durchführung und Auswertung standardisiert. Aber auch dieser Versuch hat seine Probleme. Sprache ist nur kontextbezogen verständlich, ihre Bedeutung muss von den Beteiligten stets neu ausgehandelt werden. Ohne direkte Kommunikation ist jedoch genau das nicht möglich. Eine Hilfe können strukturierte Beobachtungs- und Auswertungsbögen sein. Verbunden mit einer Schulung der BeurteilerInnen versprechen sie eine verbesserte – allerdings immer noch begrenzte – Übereinstimmung der Urteile.

### *1.3 Wie verlässlich sind verschiedene Beurteilungsverfahren? (Reliabilität)*

Eine Beurteilung soll von äußeren Umständen (Tageszeit, Reihenfolge der Prüflinge und ähnlichen Randbedingungen) unabhängig sein. Dies ist weder bei Tests noch bei informellen Leistungsproben und Beobachtungen gewährleistet.

#### *1.3.1 Die Zuverlässigkeit des Lehrerurteils*

Die Zahl von Prüfungen und die Reihenfolge, in der Leistungen bewertet werden, hat Einfluss auf die Note. Auch wenn dieselben LehrerInnen Arbeiten in zeitlichem Abstand erneut korrigieren, verändern sich ihre Bewertungen erheblich. Da zudem die Leistungen der SchülerInnen situationsabhängig variieren, müssen für Beurteilungen auf jeden Fall mehrere Leistungsproben erhoben werden – möglichst in unterschiedlichen Aufgabenformen und zu unterschiedlichen Zeitpunkten.

#### *1.3.2 Die Zuverlässigkeit von Tests*

Das Problem bei Einzelfalldiagnosen ist die breite Schwankung einer punktuell erfassten Testleistung um den ›wahren Wert‹ der eigentlich angezielten Fähigkeit (= hoher Messfehler bei Individualdaten). Bei Aussagen über größere Gruppen, wie sie für bildungspolitische Entscheidungen genutzt werden, stellt sich dieses Problem in geringerem Umfang, weil sich individuelle Schwankungen in den Kennwerten für die Stichprobe insgesamt ausgleichen. Insofern liefern Studien wie PISA, IGLU und VERA verlässliche Daten für eine schulübergreifende Systemevaluation. Ihre Ergebnisse sind aber nicht zureichend verlässlich für die Erfassung und Bewertung individueller Leistungen von SchülerInnen (oder auch LehrerInnen).

#### *1.3.3 Zwischenbilanz zu ›Reliabilität‹*

Auf der Individualebene sind sowohl Lehrerurteile als auch Tests sehr unzuverlässig. Punktuelle Leistungsproben bzw. Beobachtungen reichen deshalb in keinem Fall aus, um institutionelle Förder- oder gar Selektionsentscheidungen abzusichern. Je folgenreicher die Entscheidung für die Betroffenen, umso weniger genügt eine einzige Leistungsprobe. Außerdem sollten die Aufgaben*typen* variieren, um Zufallseffekte zu minimieren.

## *1.4 Fazit*

Gemessen an den drei Gütekriterien weisen alle Erhebungsformen Mängel auf. Diese Einsicht relativiert den Status von Bewertungen und entzieht ihnen die Grundlage für Selektionsentscheidungen. Die Diskussion hat aber auch gezeigt, dass die Gütekriterien in ihrem testtheoretischen Verständnis dem Gegenstand nicht voll gerecht werden: Menschliches Verhalten ist kontextabhängig und mehrdeutig. Ohne kognitive und emotionale Empathie kann es oft weder erklärt noch angemessen gewürdigt werden. Es kommt hinzu, dass Beschreibungen und Bewertungen für die Betroffenen nicht nur kognitiv nachvollziehbar, sondern auch sozial annehmbar sein müssen: Damit werden Standards wie Glaubwürdigkeit, Fairness und Verständlichkeit bedeutsam, die in der Diskussion über Leistungsbeurteilung noch zu wenig Beachtung finden (→ Kap. 6.5).

## 2. An welchen Maßstäben sollen Leistungen gemessen werden? (Bezugsnormen)

Die Bewertung einer Leistung kann sich an verschiedenen Maßstäben orientieren:

- Vergleich mit anderen Personen einer *Bezugsgruppe*, z. B. der Altersgruppe, des Jahrgangs einer bestimmten Schulform oder einer einzelnen Schulklasse (soziale oder kollektive Bezugsnorm);
- Feststellung, wie weit eine Leistung den in *Lernzielen* definierten Anforderungen entspricht (Sachnorm; Kriteriumsorientierung);
- Bestimmung des *Lernfortschritts*, bezogen auf die jeweiligen Voraussetzungen des einzelnen Kindes (Individualnorm; Entwicklungsorientierung).

Abhängig vom gewählten Maßstab wird dieselbe Leistung unterschiedlich bewertet. In der Praxis dominiert je nach Funktion mal der eine, mal der andere Maßstab. Für die Auswahl von BewerberInnen auf knappe Stellen in einem Betrieb oder einer Bildungseinrichtung wird man die KonkurrentInnen miteinander vergleichen; für die Zulassung zu einer Tätigkeit, deren Folgen andere betreffen, zum Beispiel beim Führerschein, macht die Überprüfung definierter (Mindest-) Anforderungen Sinn; zur Rückmeldung über Effekte des Unterrichts oder den Erfolg individueller Lerntätigkeit ist eher der Ausweis von Leistungsfortschritten angemessen.

### 2.1 Wo steht ein Schüler im Vergleich zu anderen? (kollektive Norm / Gruppenorientierung)

Bei der Vergabe von Noten dominiert in der Praxis der Bezug auf den Durchschnitt der jeweiligen Klasse. Schon die Anforderungen der gestellten Aufgaben variieren von Klasse zu Klasse, aber auch die Maßstäbe für die Beurteilung derselben Leistung unterscheiden sich. Normierte Tests fassen die Bezugsgruppe weiter, möglichst im Sinne einer für den Jahrgang repräsentativen Stichprobe. Aufgrund der unterschiedlichen Bezugspunkte kommt es trotz der (formal) gleichen Bezugsnorm zu unterschiedlichen Bewertungen. Gemessen an den Testwerten sind Noten über die Grenzen einer Klasse hinweg deshalb nicht vergleichbar. Innerhalb einer Klasse entsprechen sich die beiden Rangfolgen dagegen deutlich besser. Dies spricht für die diagnostische Kompetenz der LehrerInnen, verweist aber auf starke Unterschiede in den Maßstäben, insbesondere zwischen den Schulformen der Sekundarstufe (s. oben → 1.1.1). Zudem werden auch in der Grundschule so genannte Hauptfächer strenger bewertet als Nebenfächer – selbst in derselben Klasse.

Für das Lehrerurteil wie auch für Tests kann die Unterstellung einer Verteilung nach der Gauß'schen Normalverteilung (›Glockenkurve‹) Fehldeutungen nahe

legen[4], vor allem wenn selbst kleine Leistungsunterschiede um der Notendifferenzierung willen überbewertet werden. Ein zweites Problem stellt die Wirkung auf schwächere SchülerInnen dar: Obwohl sie Lernfortschritte machen, können diese nicht honoriert werden, da sich ihr Rangplatz dank des Lernzuwachses aller SchülerInnen in der Regel nicht verändert (›Karawaneneffekt‹). Damit sinkt ihre Lernmotivation.

## *2.2 Wo steht ein Schüler auf dem Weg zum Lernziel? (Sachnorm / Kriteriumsorientierung)*

Nach dem Beschluss der KMK (bereits aus dem Jahre 1968) sollen Leistungen mit Bezug auf definierte Ziele / Anforderungen bewertet werden. Bei der Vergabe von Noten, aber auch in vielen Verbalgutachten hat sich dieser Maßstab in der Praxis immer noch nicht durchgesetzt. Die Einführung von Standards durch die KMK-Vereinbarungen der letzten Jahre und ihre Umlegung auf Lehrplananforderungen, z. B. zum Ende der zweiten und vierten Klasse, verleiht dieser Forderung erneut Nachdruck.

Dabei stellt sich allerdings ein Problem: Lernen wird modelliert als eindimensionaler und linearer Zuwachs von Kompetenz. Diese Vereinfachung wird der Komplexität von Lernprozessen gerade in der Anfangsphase nicht gerecht: So führt zum Beispiel der Wechsel vom wortweisen Satzlesen zu einem inhaltsorientierten Textlesen einerseits zu einem wachsendem Tempo und besseren Inhaltsverständnis, aber gleichzeitig – zumindest phasenweise – auch zu mehr Verlesungen auf Wortebene. Lerngewinne lassen sich also nicht immer als bloß quantitative Reduktion von Fehlerquoten messen. Notwendig sind differenziertere Leistungsprofile, deren Ergebnisse vor dem Hintergrund von qualitativen Entwicklungsmodellen inhaltlich gedeutet werden müssen.

## *2.3 Welche Fortschritte hat ein Schüler gemacht? (individuelle Norm / Entwicklungsorientierung)*

Die Orientierung an der Individualnorm wird vor allem für Verbalgutachten gefordert. In ihnen geht es nicht nur um eine differenziertere Beschreibung des erreichten Leistungsstands, als dies durch Ziffernnoten möglich ist. Vor allem können auch die Bedingungen verdeutlicht werden, unter denen SchülerInnen die beschriebene Leistung erbracht haben. In Form eines ›Entwicklungsberichts‹ können Leistungen auf die jeweiligen Ausgangsbedingungen bezogen und damit – selbst bei gleicher Punktzahl in einem Test – als individuell unterschiedlicher Zuwachs ausgewiesen werden. Leistungsbeurteilung zielt außerdem nicht nur auf den Ausweis von erworbenen Kompetenzen (›summative‹ Bewertung). Sie

4 Es ist richtig, dass Leistungswerte in standardisierten Tests – wie viele soziale Daten – oftmals entsprechend der Glockenkurve ›normal‹ verteilt sind. Was für eine repräsentative Stichprobe gilt, muss aber nicht auf Klassenebene zutreffen.

hat auch eine wichtige Funktion für die Förderung von Lernen (›formative‹ Bewertung). Eine Orientierung an den individuellen Fortschritten hat besonders bei schwächeren SchülerInnen positive Auswirkungen auf ihre Lernmotivation, auf ihre Leistungen und ihre Selbsteinschätzung: SchülerInnen suchen die Gründe für Erfolg und Schwierigkeiten dann eher bei sich selbst, sie führen sie eher auf mangelnde Anstrengung als auf fehlende Begabung zurück und sie sind zuversichtlicher, Erfolg zu haben. Schon durch eine inhaltliche Kommentierung von Noten können Verbesserungen in diesen Dimensionen erreicht werden. ›Formativ‹ besonders nützlich sind Beurteilungen, wenn sie den SchülerInnen konkrete Hinweise für die weitere Arbeit geben.

### *2.4 Zwischenbilanz zu ›Bezugsnormen‹*

Trotz der Vorgaben der KMK (1968) dominiert bei der Notenvergabe faktisch die Gruppennorm – bezogen auf die einzelne Klasse. Aber auch für Verbalgutachten spielt sie neben der Kriteriumsorientierung eine wichtige Rolle: Die individuelle Bezugsnorm kommt nur in einer Minderheit von etwa 10% der Aussagen zur Geltung. Grundsätzlich hat eine Rückmeldung von individuellen Lernfortschritten (statt einer Bewertung von Leistungen im Vergleich mit einer Bezugsgruppe) positivere Effekte auf leistungsschwächere SchülerInnen. Aber je nach Funktion haben auch die Zielorientierung und der Gruppenvergleich ihre Berechtigung – als ergänzende Information (vgl. unten → Kap. 6.4).

## *3. Wie werden verschiedene Formen der Leistungsbeurteilung umgesetzt, und welche Wirkungen haben sie?*

In → Kap. 2 wurden die Effekte unterschiedlicher Bezugsnormen *grundsätzlich* untersucht. Im Folgenden werden ihre Umsetzung in der Praxis und damit Unterschiede zwischen den im *Schulalltag* gebräuchlichen Formen der Bewertung analysiert. Wie sich zeigen wird, greift die Gleichsetzung ›Ziffernnoten = soziale Bezugsnorm‹ bzw. ›Verbalgutachten = individuelle Bezugsnorm‹ zu kurz.

### *3.1 Wie weit werden Ziffernoten und Verbalgutachten ihren eigenen Ansprüchen gerecht?*

*Ziffernnoten* werden vor allem zwei Vorteile unterstellt: Verständlichkeit und Vergleichbarkeit. Diese Erwartungen (→ vgl. unten 4.) werden allerdings kaum erfüllt, wie die in → Kap. 1 und 2 resümierten Untersuchungen gezeigt haben. Die Vergleichbarkeit wird allenfalls im Klassenrahmen erreicht, und die Pauschalität der Ziffer wird den unterschiedlichen Teilleistungen innerhalb eines Faches nicht gerecht. Noten wird zu Recht vorgeworfen, dass dieselbe Ziffer sehr Unterschiedliches bedeuten kann: zum Beispiel können die Nachlässigkeit einer kompetenten Schülerin und die große Anstrengung eines schwachen Schülers zu formal gleichen Leistungen führen, die dann auch gleich benotet werden. In-

sofern ist eine sprachliche Beschreibung aussagekräftiger. Eine Differenzierung nach Teilkriterien führt zusätzlich – bei jeder Form der Beurteilung – zu valideren Aussagen. Insofern ist bereits die traditionelle Auffächerung etwa der Aufsatznote in ›Inhalt / Sprache / Form‹ oder der Gesamtnote für Deutsch in ›Sprachgebrauch / Lesen / Rechtschreiben‹ (wie in NRW) aussagekräftiger. Im Vergleich zu den Möglichkeiten einer Verbalbeurteilung bleibt aber selbst diese Differenzierung noch zu grob, um die Komplexität der angestrebten Fähigkeiten angemessen zu erfassen.

Weitere Einschränkungen bei der Umsetzung sind bereits in den vorhergehenden Kapiteln diskutiert worden. Sie betreffen bei den Ziffernnoten vor allem die Dominanz der Gruppennorm, bei den *Verbalgutachten* die Vernachlässigung der Entwicklungsnorm. Denn auch Verbalbeurteilungen erfüllen die in sie gesetzten Erwartungen meist nicht. Die empirischen Studien zeigen folgende Mängel:

- fehlender Bezug auf die individuelle Leistungsentwicklung;
- Ungleichgewicht der Fächer und Leistungsdimensionen, d. h. starke Dominanz der Lese-, Rechtschreib- und Rechenleistungen;
- fehlende Fördervorschläge;
- Beschönigung der Rückmeldungen;
- Standardisierung der Aussagen durch Nutzung von Textbausteinen.

Diese Kritik wurde bereits in den ersten Evaluationen Anfang und Mitte der 1980er Jahre geäußert, findet sich aber in unveränderter Form bis heute. Vor diesem Hintergrund werden sich die grundsätzlich positiven Effekte einer entwicklungsorientierten Bewertung von Leistungen (→ Kap. 2.3) im Schulalltag vermutlich kaum wiederfinden.

## *3.2 Welche (Neben-)Wirkungen haben verschiedene Beurteilungsformen?*

Im Folgenden wird in mehreren Schritten untersucht, ob und ggf. wie veränderte Beurteilungsformen überhaupt den Unterricht und daraufhin die Motivation, die Leistung und das Selbstkonzept von SchülerInnen verändern.

### *3.2.1 Gibt es einen Zusammenhang zwischen Unterrichtskonzept und Beurteilungsform?*

Es besteht eine wechselseitige Abhängigkeit zwischen dem Unterrichtsstil der Lehrperson und der Form der Leistungsbeurteilung. LehrerInnen, die sich an der Individualnorm orientieren, differenzieren auch die Lernangebote stärker. Umgekehrt zeigt die Praxis aller Reformschulen, dass sich Unterricht nur verändern lässt, wenn die Verfahren der Leistungsbewertung auf den didaktischen Ansatz abgestimmt werden.

Für die Einführung neuer Formen der Leistungsbeurteilung stellt deshalb die weithin noch unveränderte Unterrichtskultur ein besonderes Problem dar. Insofern verwundert die Häufigkeit der in → Kap. 3.1 berichteten Fehlformen ver-

baler Beurteilungen nicht. Ein individualisierender Unterricht mit individuellen Rückmeldungen ist immer noch nicht sehr verbreitet. Sofern seine Prinzipien umgesetzt werden, geschieht dies zudem meist mit starken inhaltlichen Einschränkungen. Offener Unterricht, der über eine organisatorische Differenzierung ›von oben‹ hinausgeht, ist auch in Grundschulen die Ausnahme. Die Einführung pädagogisch anspruchsvollerer Formen der Leistungsbewertung verlangt also eine umfassendere Reform des Unterrichts. Umgekehrt kann aber auch die Einführung oder zumindest das Zulassen differenzierterer Formen der Leistungsbeurteilung Räume öffnen für ebensolche Initiativen. Im Blick auf die anspruchsvoll formulierten Ziele der eigenen Richtlinien haben Kultusministerien hier eine Verantwortung, die sie nicht an zufällige Eltern-/Lehrer-Mehrheiten in den Schulkonferenzen abtreten dürfen.

*3.2.2 Beeinflusst die gewählte Beurteilungsform das Unterrichtsklima?*

Schulfreude variiert mit der durch Noten definierten Leistungsposition in der Klasse. Sie nimmt generell mit dem Alter – und damit auch wachsender Schul- und Bewertungserfahrung – ab. In Klassen ohne Noten fühlen sich Kinder wohler. Entgegen häufigen Vorurteilen beeinträchtigt dies nicht die Leistungsbereitschaft. Allerdings zeigen die Kinder im Durchschnitt auch keine besseren Leistungen. Eine Erhöhung des Selektionsdrucks verschärft jedoch die Konkurrenz und führt im unteren Leistungssegment zu einem Gegeneinander von SchülerInnen und LehrerInnen, wie vor allem Studien zum high-stakes testing in den USA zeigen. Dort hängen nicht nur die Schulkarrieren der SchülerInnen, sondern auch die Gehälter der LehrerInnen und finanzielle Zuweisungen an Schulen von den Ergebnissen in Vergleichstests ab. Dies führt u. a. zu höheren drop-out-(treffender: push-out-)Quoten und generell zu schlechteren Ergebnissen im unteren Leistungsbereich und in den Gruppen der gesellschaftlichen Minderheiten.

*3.2.3 Beeinflusst die gewählte Beurteilungsform zentrale Merkmale der Persönlichkeitsentwicklung?*

Ob Kinder Leistungen erbringen, hängt nicht nur von ihren kognitiven Grundfähigkeiten, sondern in starkem Maße auch von Motivation und Emotionen ab. Aber nicht nur wegen dieser ›instrumentellen‹ Bedeutung, sondern auch wegen ihres Eigenwerts ist zu prüfen, wie die Art der Leistungsbeurteilung die Persönlichkeitsentwicklung beeinflusst. Denn die Schule dient primär nicht der Stabilisierung der Gesellschaft, sondern der Entwicklung des individuellen Potenzials der Kinder. Vor allem die Grundschule hat nicht nur einen Unterrichts-, sondern auch einen Erziehungsauftrag.

*3.2.3.1 Beeinträchtigen oder stützen Ziffernnoten bzw. Verbalgutachten die Lernmotivation?*

Wie die in → Kap. 2 resümierten Grundlagenstudien gezeigt haben, wirkt sich eine Entwicklungsorientierung der Leistungsbeurteilung positiv auf die Lern- und Leistungsmotivation aus – positiver als Bewertungen mit Bezug auf die Position in der Vergleichsgruppe. Beim Vergleich von Ziffernnoten und Verbalbeurteilungen im Unterrichtsalltag kann dieser Vorteil zwar nicht generell bestätigt

werden. Allerdings profitieren schwächere und ängstliche SchülerInnen auch hier von einem Notenverzicht. Bei leistungsstärkeren Kindern wiederum wird eine stärkere Orientierung an äußerer Anerkennung beobachtet (›externale Motivation‹) – auf Kosten persönlicher Interessen und damit einer *intrinsischen* Motivation. In Schulversuchen, die auch nach einem anderen pädagogischen Konzept arbeiten, fallen die Ergebnisse meist positiver aus als bei breiten Erhebungen in Schulen, die noch in traditionellen Formen oder unter unveränderten Rahmenbedingungen arbeiten.

#### *3.2.3.2 Verringern oder vergrößern Ziffernnoten bzw. Verbalgutachten die Schul- und Prüfungsangst?*

Zeugnisse und Noten zählen für Kinder zu den stärksten Angstauslösern. Dies belegen sehr konkret Statistiken der Kinder- und Jugendtelefone. Zwar werden Klassenarbeiten ohne Noten nur von einem Teil der Kinder als weniger Angst auslösend eingeschätzt. Differenziertere Befragungen zeigen aber, wie sehr die Sorge um Noten Kinder bewegt – im Vergleich sogar stärker als die Angst vor Kriegsgefahren. Dabei verdoppelt sich der Anteil der betroffenen Kinder vom Ende der Grundschulzeit bis in die Sekundarstufe hinein. Wiederum sind die Kinder im unteren Leistungssegment besonders betroffen – in der Sekundarstufe aber auch Kinder mit guten Noten.

Noten haben zwar in verschiedenen Bundesländern ein unterschiedliches Gewicht für den Zugang zu den Schulformen der Sekundarstufe; diese institutionellen Differenzen beeinflussen die Prüfungsangst aber deutlich weniger als die Art und Weise, wie LehrerInnen innerhalb des jeweiligen Systems mit den Noten umgehen. Insgesamt scheint das ›Notenklima‹ in einer Klasse von besonderer Bedeutung für die emotionalen Effekte zu sein. Auch hier ist das Ausmaß des Selektionsdrucks bedeutsamer als die Form der Beurteilung (Ziffer oder verbale Beschreibung).

#### *3.2.3.3 Schädigen oder stärken Ziffernnoten bzw. Verbalgutachten das Selbstkonzept?*

Das Verhältnis von Selbstwertgefühl und Leistung ist wechselseitig: Wer sich etwas zutraut, erbringt bessere Leistungen. Positive Rückmeldungen zur eigenen Leistung wiederum steigern das Selbstwertgefühl von SchülerInnen. Dies bestätigen die in → Kap. 2 referierten Studien. Aber auch unter diesem Aspekt zeigen die Vergleiche von Ziffernnoten und aktuell praktizierten Formen der Verbalbeurteilung im Schulalltag kaum Unterschiede – vermutlich bedingt durch die unzureichende Umsetzung der Individualnorm in Gutachtenzeugnissen.

### *3.2.4 Belasten oder fördern Ziffernnoten bzw. Verbalgutachten die Leistungsentwicklung?*

Entgegen den vielerorts geäußerten Befürchtungen führt ein Verzicht auf Noten nicht zu einem Leistungsabfall. Zugleich zeigen die Studien unter Alltagsbedingungen, dass zumindest die gegenwärtig üblichen Formen von Verbalgutachten nicht zu einer Verbesserung der Schülerleistungen führen. Allerdings ist nach

Studien aus den angelsächsischen Ländern zu befürchten, dass eine Verschärfung des Notendrucks zu einer Verschlechterung von Bildungschancen und Leistungen im unteren Leistungssegment und bei gesellschaftlichen Minderheiten führen. Belege für eine Verbesserung von Leistungen durch die Vergabe von Noten fehlen ganz. Die verbreitete Annahme, dass SchülerInnen nur lernen, wenn ein entsprechender Druck ausgeübt wird, ist falsch. Auch auf Systemebene zeigen die internationalen Studien, dass Länder mit früher Benotung von Leistungen oder mit externen Prüfungen im Leistungsvergleich nicht besser abschneiden als Länder ohne diese Merkmale.

#### *3.2.5 Zwischenbilanz zu ›Wirkungen‹*

Über alle Untersuchungen hinweg finden sich nur wenige und zudem in der Regel nur schwach ausgeprägte Unterschiede in den Effekten. Diese Befunde aus dem Schulalltag irritieren nach den in → Kap. 2 berichteten positiven Vorteilen differenzierterer Beurteilungsformen in kontrollierten Versuchen. Das Ergebnis lässt sich am ehesten auf den folgenden Nenner bringen: Grundsätzlich kann eine Veränderung der Bewertungsformen positive Effekte haben. Diese werden aber im Schulalltag nur gebremst wirksam – sei es, dass Mischformen (z. B. Ziffernnoten kombiniert mit erläuterndem Bericht) verwendet werden, sei es, dass die Intentionen der Verbalgutachten in Inhalt und Form nicht (zureichend) umgesetzt werden oder der externe Selektionsdruck das Potenzial einer förderorientierten Rückmeldung nicht zur Entfaltung kommen lässt.

## *4. Wie gut erfüllen Ziffernnoten und Verbalgutachten wichtige Funktionen aus der Sicht der Betroffenen?*

Mit der Beurteilung von Leistungen werden verschiedene Erwartungen verbunden (→ Kap. 0.3). Wie in den → Kap. 1 – 3 gezeigt, können Noten diese Ansprüche nicht erfüllen. Aber auch Verbalgutachten werden den gesetzten Anforderungen im Schulalltag meist nicht gerecht. Für die Entscheidung, ob Noten abgeschafft und durch Verbalbeurteilungen ersetzt werden sollen (und können …), ist deshalb wichtig zu wissen, wie die Betroffenen Vor- und Nachteile der verschiedenen Beurteilungsformen wahrnehmen.

### *4.1 Einschätzungen von LehrerInnen*

Anfang der 1970er Jahre hielten fast 80% der LehrerInnen Noten für unverzichtbar, 2005 ist es nur noch gut die Hälfte. Dabei gibt es große Unterschiede zwischen den Schularten. In den Grund- und Sonderschulen halten inzwischen fast zwei Drittel der LehrerInnen Noten für überflüssig, in den Schulformen der Sekundarstufe sind es Minderheiten von nur etwa einem Viertel. Diese Schultypen haben allerdings auch weniger Erfahrung mit Verbalgutachten – und die externen Selektionserwartungen haben ein stärkeres Gewicht.

Für Noten sprechen aus Sicht der LehrerInnen die Erwartungen von SchülerInnen und Eltern, während sie selbst viele Vorteile in der Verbalbeurteilung sehen. Aber auch die BefürworterInnen beklagen den erheblichen Mehraufwand an Arbeit, den eine kontinuierliche Beobachtung und ihre sprachliche Würdigung verlangen. Dabei müsste zumindest die Sammlung und Sichtung differenzierter Daten Grundlage einer *jeden* Leistungsbeurteilung sein.

## *4.2 Einschätzungen von SchülerInnen*

Die Befundlage ist widersprüchlich. Forschungsmethodisch liegt das an der unterschiedlichen Art zu fragen, inhaltlich aber auch an der unterschiedlich stark ausgeprägten Erfahrung mit Alternativen zu den herkömmlichen Noten. Einerseits zeigen viele Daten, dass Noten von SchülerInnen (vor allem im unteren Leistungssegment) als sehr belastend erlebt werden. Andererseits überraschen immer wieder die Ergebnisse von Umfragen, wonach die Mehrheit der SchülerInnen an Noten festhalten wolle. Allerdings gilt auch dies eher für leistungsstarke als für leistungsschwache SchülerInnen. Stärker als bei Eltern und LehrerInnen ist das Meinungsbild polarisiert. Zudem finden Verbalbeurteilungen in den ersten Schuljahren mehr Zustimmung, während ältere SchülerInnen mehrheitlich für Ziffernnoten plädieren (besonders stark in der Hauptschule, weniger deutlich im Gymnasium).

## *4.3 Einschätzungen von Eltern*

In den 1970er Jahren sprachen sich drei Viertel der Eltern für Noten aus. Auch heute plädiert nur eine Minderheit für ihre Abschaffung (zudem mit wieder leicht abnehmender Tendenz in den letzten Jahren). In den ersten Schuljahren findet das Verbalzeugnis noch eine hohe Zustimmung, die aber zum Ende der Grundschulzeit auf etwa ein Viertel absinkt. Allerdings plädieren Eltern aus Schulen mit breiter etablierter Verbalbeurteilung (in einzelnen Fächern / generell bis Klasse 3/4) für deren Beibehaltung. Insgesamt neigen Eltern in ihrer Mehrheit zu einer Kombination von Ziffernnoten mit verbalen Kommentaren, wobei das Ziffernzeugnis – anders als bei den Kindern – die stärkste Zustimmung unter den (zukünftigen) Gymnasialeltern findet.

## *4.4 Einschätzungen von Arbeitgebern*

Befragte Unternehmen gewichten die Eindrücke aus dem Einstellungsgespräch für die Auswahl von BewerberInnen deutlich höher als die Noten im Zeugnis. Auch die Aussagekraft von Berufsschul- und IHK-Zeugnissen für die spätere berufliche Bewährung wird gering eingeschätzt. Dass größere Unternehmen und viele Kammern regelmäßig eigene Leistungstests durchführen, bestätigt die Ergebnisse aus Befragungen: Arbeitgeber trauen den Noten – vor allem unter dem Gesichtspunkt der Vergleichbarkeit – nicht. Auch unternehmensintern haben Mitarbeitergespräche und lernbiografische Dokumente an Einfluss für die Leistungsbewertung gewonnen.

## 4.5 Einschätzungen in der Öffentlichkeit

In der Öffentlichkeit herrscht generell eine konservative Haltung vor. Wie beim Sitzenbleiben und bei den Hausaufgaben sprechen sich auch bei den ›Schulnoten im üblichen Sinn‹ nur 30–40% der Befragten für deren Abschaffung aus. Dieser Befund passt zu deutschen und US-amerikanischen Daten, wonach Personen (ungewohnten) Praktiken in der Schule umso kritischer gegenüber stehen, je weniger Kontakt sie zur Schule (über eigene Kinder oder Enkel) haben.

## 4.6 Zwischenbilanz zu den ›Einschätzungen‹ durch die Beteiligten

Eine Abschaffung von Noten findet in Befragungen kaum Zustimmung. Dabei stehen die gegebenen Begründungen (Vergleichbarkeit, Eindeutigkeit) oft in explizitem Widerspruch zu den empirischen Befunden, sind also sachlich nicht gerechtfertigt.

Dieses Ergebnis muss aber differenziert werden. Befragt nach einzelnen Punkten (›macht mir Angst‹, ›nimmt meinem Kind die Motivation‹) äußern sich Mehrheiten in den einzelnen Teilgruppen oft zensurenkritisch. Für Kinder wie LehrerInnen spielt das indirekte Argument der höheren Akzeptanz der Noten bei Eltern, Verwandten und ›Abnehmern‹ eine wichtige Rolle für die eigene Zustimmung. Arbeitgeber selbst dagegen verlassen sich bei der Auswahl von BewerberInnen nur sehr eingeschränkt auf Noten in Abschlusszeugnissen.

Insgesamt lassen sich zwei Trends beobachten:

- Personen, die (länger) Erfahrungen mit Verbalgutachten haben, äußern sich generell positiver zu dieser Form der Beurteilung.
- Vor allem die Eltern tendieren zu einer Verbindung von Ziffern und verbalen Aussagen.

Nimmt man die in → Kap. 1 bis 3 referierte Breite erdrückender Sachargumente und betrachtet man andererseits, wie langsam sie in der Breite wahrgenommen werden, so stellt sich die Frage, ob hier nicht der Gesetzgeber gefordert ist. So wichtig Mitbestimmungsrechte der Betroffenen für die Gestaltung des Schulalltags sind – grundsätzliche Entscheidungen wie die Beurteilung von Leistungen sind gesamtgesellschaftlich zu verantworten, solange sie so nachhaltige Konsequenzen haben wie in unserem selektiven System (→ Kap. 7).

## 5. Rechtfertigt der Ertrag aufwändigere Formen der Erhebung und Bewertung von Leistungen?

Studien zur Arbeitsbelastung von LehrerInnen unterscheiden den zeitlichen Aufwand von der psychischen Belastung. Sie zeigen:

- Teiltätigkeiten der Leistungsbeurteilung gehören generell zu den *zeitlich* aufwändigeren Aufgaben;
- das Benoten ist nach Meinung der LehrerInnen ganz erheblich weniger aufwändig als das Schreiben von Entwicklungsberichten, und dieses ist auch deutlich aufwändiger als das Korrigieren von Arbeiten oder das Schreiben von Zeugnissen.

Betrachtet man die *psychische* Belastung, so wird das sehr zeitaufwändige Korrigieren als nur wenig anstrengend empfunden, während alle Formen der Bewertung von Leistung zu den fünf belastendsten Tätigkeiten zählen. Unter ihnen rangieren die Entwicklungsberichte allerdings wieder eindeutig auf Platz 1, während Noten und Zeugnisse als weniger belastend eingeschätzt werden.

Die beiden Vergleiche machen deutlich, dass entwicklungsorientierte Lernberichte nicht nur einen hohen zeitlichen Aufwand erfordern, sondern von den LehrerInnen auch als besondere Belastung empfunden werden. Einer Person sprachlich differenziert gerecht zu werden, sich nicht hinter einer (scheinbar objektiveren) Verrechnung von Daten verstecken zu können, ist offensichtlich eine hohe Anforderung – vor allem, wenn diese Beurteilungen selektionswirksam werden. So berechtigt die Anforderung einer differenzierten Rückmeldung nach der Kritik an Noten auch sein mag – ohne entsprechende Ausbildung und Unterstützung (Austausch, Supervision) ist sie in der Breite wohl nicht erfüllbar.

Wer fordert, dass Ziffernnoten und -zeugnisse durch gehaltvolle Entwicklungsberichte ersetzt werden, muss LehrerInnen also zeigen, welchen Vorteil sie von dieser zusätzlichen Anforderung haben, oder ihre zeitliche Mehrbelastung durch eine Gratifikation ausgleichen – sonst wird der Anspruch der Reform unterlaufen, wie die ernüchternden Ergebnisse aus Inhaltsanalysen von Verbalgutachten zeigen (→ Kap. 3.1). Ermutigend sind die Erfahrungen aus Reformschulen und aus größeren Schulversuchen, wonach LehrerInnen ihre pädagogische Arbeit als sinnvoller erleben und zufriedener im Beruf sind, wenn differenziertere Formen der Rückmeldung gemeinsam eingeführt werden.

Beachtung verdient aber auch der Befund, dass alle Formen der Leistungsbeurteilung als besonders belastend wahrgenommen werden, wenn sie der Selektion dienen. Hier wird ein grundsätzlicheres Problem deutlich, das über technische Fragen der Darstellungsform hinausweist (→ Kap. 7).

# 6. Zwischenbilanz und pädagogische Folgerungen

Keines der diskutierten Verfahren, Leistungen zu *erfassen*, ist – für sich genommen – valide, objektiv und verlässlich genug, um Einzelfallentscheidungen über Bildungskarrieren zu rechtfertigen. Sowohl punktuelle Tests als auch informelle Beobachtungen sind fehleranfällig, methodische Verbesserungen sind nur begrenzt möglich (→ Kap. 1).

Als Maßstab für *Bewertungen* reicht keine der diskutierten Bezugsnormen allein aus, um die Informationsbedürfnisse der verschiedenen Zielgruppen von Beurteilungen zu befriedigen. Leistungsbewertungen haben zu unterschiedliche Funktionen, als dass eine einzige Form sie erfüllen könnte (→ Kap. 2).

Die *Wirkungen* sowohl von Ziffernnoten wie auch von Verbalzeugnissen sind aus pädagogischer Sicht kritisch einzuschätzen. Dies hängt – vor allem bei den Ziffernnoten – mit den oben genannten methodischen Problemen zusammen, andererseits – vor allem bei den Verbalgutachten – mit Unzulänglichkeiten ihrer Umsetzung im Schulalltag (→ Kap. 3). Ein besonderes Problem stellt für beide die starke Selektionsorientierung der Schule dar (vgl. dazu unten → Kap. 7).

## 6.1 Grundlegende Einwände

Vor der weiteren Diskussion alternativer *Formen* der Bewertung sind drei grundsätzliche Fragen zu beantworten:

### *6.1.1 Genereller Verzicht auf eine Rückmeldung zu Leistungen?*

Rückmeldungen zu Leistungen sind unverzichtbar, wenn Lernende sich entwickeln sollen. Sie sind lernförderlich, wenn sie sachbezogen erfolgen, d. h. individuelle Stärken und Schwächen benennen und vor allem Hinweise für konkrete Lern- und Fördermöglichkeiten geben: *Beschreibung* statt *Bewertung*. Im Sinne von von Hentigs bekannter Forderung »Die Menschen stärken, die Sachen klären«, setzen sie aber auch Sensibilität für mögliche Nebenwirkungen auf die Betroffenen voraus. Ohne Anerkennung der individuellen Fortschritte und ohne Hilfen für die Überwindung von Schwächen sind die – auch notwendigen – Hinweise auf Fehler kontraproduktiv.

### *6.1.2 Verzicht auf eine Zertifizierung nach außen?*

Auf Zertifizierung von Leistungen nach außen und damit auf formalisierte Bewertungen zu verzichten ist ambivalent. Der Verzicht auf eine Zertifizierung nach außen entlastet den Unterricht, würde aber zu einer Zunahme externer Prüfungen mit all ihren Nachteilen führen. Für die Leistungsbewertung wäre vor allem die Konsequenz, dass Leistungsproben sich auf eine punktuelle Bewährungssituation konzentrieren, problematisch. Der Unterricht wiederum geriete über externe Prüfungen in Abhängigkeit von fremd bestimmten Anforderungen, die dann die pädagogisch begründeten Lernziele überlagern.

*6.1.3 Verzicht auf Ziffernoten als Form der Beurteilung?*

Bleibt die Frage, ob Beurteilungen weiterhin in Form von Noten erfolgen sollen,

- deren Grundlage informelle Leistungsnachweise sind,
- die mit Bezug auf den Rang in der Klasse bewertet und
- die in Form von Ziffern dargestellt werden sowie
- Selektionsentscheidungen begründen sollen.

Die in diesem Gutachten referierten Studien stellen Noten unter allen vier Gesichtspunkten in Frage: Die Bewertung nach informellen Proben und Beobachtungen ist in hohem Maße fehleranfällig, die soziale Bezugsnorm hat negative Auswirkungen auf die Lernmotivation, und vor allem sind Ziffern wenig aussagekräftig, suggerieren zudem eine Genauigkeit, Vergleichbarkeit und Prognosefähigkeit, die sie nicht gewährleisten können.

Bleibt die Frage nach den Alternativen, zumal auch Verbalgutachten die in sie gesetzten Erwartungen bisher nicht erfüllt haben.

## *6.2 Keine Beurteilungsform erfüllt alle Anforderungen – einfache Auswege aus dem Bewertungsdilemma gibt es nicht*

Leistungsbeurteilungen haben unterschiedliche Funktionen zu erfüllen. Je nachdem, ob die Förder-, Berichts- oder Selektionsfunktion im Vordergrund steht, und je nach den Adressaten sind verschiedene Formen angemessen.

Zumindest in der Grundschule kann (und sollte) schon heute auf Noten verzichtet werden. Vorrang hat eine möglichst differenzierte Rückmeldung der individuellen Leistung und ihrer Entwicklung. Verbalbeurteilungen dürfen dabei aber nicht bloße Übersetzungen der Ziffernnoten sein, da sie dann eine überflüssige Zusatzbelastung sind. Sie müssen vielmehr durch die Differenzierung von Teilleistungen und durch den Bezug der Bewertung auf die individuelle Entwicklung die Notenbewertung inhaltlich ergänzen.

Damit die Aussagekraft von Beurteilungen generell verbessert werden kann, sind folgende Anforderungen zu berücksichtigen:

- Für die Beurteilung sind *Kriterien* zu entwickeln, die sich auf die Ziele des Unterrichts und nicht nur auf spezifische Aufgaben beziehen. So kann LehrerInnen geholfen werden, ein tieferes Verständnis der Ziele von Unterricht zu gewinnen und die Beurteilung besser auf diese abzustimmen.
- LehrerInnen brauchen mehr *Aus- und Fortbildung*, die sie für die Risiken der Leistungsbewertung sensibilisiert und auf ihre unterschiedlichen Funktionen vorbereitet. Erfolgreich sind solche Maßnahmen, wenn sie sich auf konkrete Beispiele einer good-practice beziehen können.
- Eine unterrichtsbegleitende Abstimmung von Kriterien im *Austausch* über konkrete Bewertungsversuche hilft LehrerInnen, Klarheit über die Ziele von Unterricht und darauf bezogene Beurteilungskriterien zu gewinnen.

Unter dem Gesichtspunkt der Akzeptanz in der Praxis scheint zurzeit allerdings eine Kombination von Ziffern und verbalen Kommentaren als Zwischenschritt am ehesten Erfolg zu versprechen.

## *6.3 Daten aus verschiedenen Erhebungsverfahren sind miteinander zu verbinden*

Wo immer möglich – und bei der Nutzung für Selektionsentscheidungen zwingend – sind Leistungsdaten

- zu mehreren Zeitpunkten
- anhand verschiedener Aufgaben
- in unterschiedlichen Situationen

zu erheben.

In diesem Rahmen sollten normierte Tests und strukturierte Beobachtungs- und Bewertungsraster eine stärkere Rolle spielen als bisher üblich. Hier hat das zentrale ›Institut für Qualität im Bildungswesen‹ (IQB) in Berlin einige Entwicklungsarbeit zu leisten, um wenigstens das Niveau zu erreichen, auf dem etwa das niederländische CITO-Institut Schulen Instrumente zur Evaluation von Unterricht anbietet (nicht vorschreibt!). Das setzt voraus, dass Fachdidaktik und Unterrichtspraxis gewichtig an der Entwicklung von Aufgaben beteiligt werden. Selbst dann bleibt zu bedenken, dass Tests nur bestimmte Leistungstypen erfassen können und dass ihre Ergebnisse generell interpretationsbedürftig sind: Zahlen sprechen nicht für sich. Insofern können Tests informelle Beobachtungen und vor allem das Lehrerurteil selbst nicht ersetzen.

## *6.4 Bewertungen müssen auf unterschiedliche Bezugsnormen bezogen werden*

Je nach der Funktion von Beurteilungen sind unterschiedliche Kriterien angemessen. Um Missverständnisse auszuschalten ist der Maßstab einer Bewertung explizit zu machen. In der Schule müssen die Förder- und Berichtsfunktion Vorrang haben. Beurteilungen sollten sich deshalb auf die Entwicklungs- und Sachnorm (Lernziel) beziehen. Auch für Selektionsentscheidungen reicht in der Regel der Nachweis (noch) fehlender Kompetenzen aus. Sollte die soziale Bezugsnorm tatsächlich wichtig werden (z. B. bei fehlenden Plätzen in Ausbildungsgängen), sind Noten durch Prozentrangangaben (z. B. zu Leistungen in normierten Tests) zu ersetzen.

## *6.5 In dialogischer Form sollten Fremd- durch Selbstbeurteilungen ergänzt werden*

Welche Formen der Leistungsbeurteilung praktiziert werden, ist eng verknüpft mit dem Bild, das LehrerInnen von Kindern, von ihrer Lernfähigkeit und von der eigenen Rolle haben. In den bisherigen Überlegungen ist Leistungsbewertung

– wie traditionell in der Schule üblich – überwiegend als Urteil von Lehrenden über Lernende verstanden worden. Diese einseitige Abhängigkeit ist nicht mit der begrenzten Aussagekraft und hohen Fehleranfälligkeit von Fremdurteilen zu vereinbaren. Sie widerspricht zudem dem Anspruch einer demokratischen Schule, junge Menschen in der Entwicklung ihrer Selbstständigkeit zu unterstützen. Diese setzt die Bereitschaft und Fähigkeit voraus, sich selbst und insbesondere konkrete Leistungen kritisch einzuschätzen. Beurteilung sollte in der Schule – wie zum Teil auch schon im Berufsleben üblich – deshalb als dialogischer Prozess ausgestaltet werden. Die Kompetenz der Selbsteinschätzung zu fördern müsste ein zentrales Ziel von Unterricht sein. Empirische Studien zeigen außerdem, dass kontroll-orientierte Rückmeldungen nicht nur die Motivation, sondern auch die Leistung beeinträchtigen. Selbst gute Noten wirken auf Dauer kontraproduktiv, weil sie die intrinsische Motivation beeinträchtigen.

Verbalbeurteilungen wird zu Recht ihre Subjektivität vorgehalten. Zugleich ist dies aber auch ihre Stärke, da sie – anders als Ziffernnoten – keine Schein-Objektivität vorgaukeln. Verbale Beurteilungen geben zurzeit zwar eher Auskunft über die Beurteilungspraxis der Lehrkraft; sie machen damit aber das Dilemma der Beurteilungspraxis offenkundig und zwingen zur Suche nach Verfahren, um die unvermeidliche Subjektivität zu kontrollieren. Leistungsbeurteilung ist ein kommunikativer Akt. Seine Mehrdeutigkeit kann nicht technisch-methodisch ausgeschaltet, sondern nur sozial kontrolliert werden. Je nachhaltiger die Folgen einer Beurteilung, umso wichtiger wird deshalb die Beteiligung verschiedener Personen – vor allem der Betroffenen selbst.

## *7. Fazit und bildungspolitische Bewertung*

Unser Fazit zur Ausgangsfrage ›Sind Noten nützlich – und nötig?‹ fällt negativ aus.

Zum einen erfüllen Noten die Erwartungen ihrer Befürworter nicht:

- Sie sind nicht valider, objektiver und zuverlässiger als andere Beurteilungsformen.
- Die beanspruchte Vergleichbarkeit ist durch den in der Regel üblichen Bezug auf den Klassendurchschnitt und die unvermeidlichen Beurteilungsfehler sehr eingeschränkt.
- Ziffernnoten erfüllen die verschiedenen Funktionen der Leistungsbeurteilung (Motivation, Information) nicht besser, zum Teil sogar schlechter als andere Formen der Rückmeldung.

Wenn Noten im Schulalltag trotzdem so viel Zustimmung finden, so hängt dies vermutlich damit zusammen, dass sie SchülerInnen und Eltern vertraut sind. Für LehrerInnen ist ihre Vergabe außerdem mit einem geringeren Arbeitsaufwand verbunden als das Schreiben von Verbalgutachten. Schließlich suggeriert ihre nur

scheinbare[5] Verrechenbarkeit eine Vereinfachung von Selektionsentscheidungen. Diese haben im deutschen Schulsystem eine hohe – und im Vergleich zu anderen Ländern erheblich höhere – Bedeutung. Damit ist der institutionelle Kontext der Leistungsbeurteilung angesprochen.

Am Ende der Grundschulzeit besuchen nur noch rund 80% der SchülerInnen eine Klasse ihres Einschulungsjahrgangs und unter den 15-Jährigen sind es kaum mehr als 60%, die eine ›glatte‹ Schullaufbahn aufweisen können. Fast 40% der SchülerInnen haben also mindestens eine der folgenden Maßnahmen erlebt: Zurückstellung am Schulanfang; Nichtversetzung; Überweisung in die Sonderschule; ›Abschulung‹ in eine niedrigere Schulform. Die Konsequenz sind kumulative Misserfolgserfahrungen – motivational die ungünstigste Voraussetzung für Leistungsbemühungen und für Lernerfolge.

Kinder aus anregungsarmen Elternhäusern sind davon besonders betroffen, da ihre soziale Herkunft, verbunden mit der selektiven Struktur des Schulsystems, sie mehrfach benachteiligt:

- Je höher der sozio-ökonomische Status der Eltern ist, umso anregungsreicher sind die Lernmöglichkeiten ihrer Kinder vor der Schule, so dass sie bessere kognitive Voraussetzungen in die Schule mitbringen.
- In sozial homogenen Stadtvierteln kommen sie in der Regel auch in eine Lerngruppe, die durch die Herkunft der anderen Kinder ebenfalls ein anregenderes Milieu bietet. Deshalb entwickeln sich auch ihre Leistungen über die Grundschulzeit hinweg besser – und damit ihre Chancen auf den Besuch einer höheren Schulform in der Sekundarstufe.
- Selbst wenn Kinder am Ende der Grundschulzeit vergleichbare Leistungen erreichen, ist ihr Zugang zu einer höheren Schulform umso wahrscheinlicher, je höher der soziale Status der Eltern ist: Sie erhalten häufiger eine Empfehlung für das Gymnasium und ihre Eltern folgen einer solchen Empfehlung auch eher. Diese Entscheidung ist deshalb bedeutsam, weil sich die Leistungen in der Sekundarstufe auch bei gleichen kognitiven Voraussetzungen und gleichem sozialen Status der Eltern umso besser entwickeln, je höher die besuchte Schulform ist.
- Aber auch wenn Kinder mit vergleichbaren Grundschulleistungen in dieselbe Schulform wechseln, fällt der Lernerfolg innerhalb dieser Schulform umso besser aus, je höher der sozio-ökonomische Status der Eltern ist, da sie u. a. ihre Kinder besser unterstützen können.

Gesteuert werden die innerschulischen Ausleseprozesse durch Noten. Diese sind aber nicht in der Lage, unterschiedliche Fähigkeiten zureichend genau auszuweisen. Beurteilte Leistungen werden überlagert durch andere Faktoren, vor allem

5 So ist es statistisch nicht zulässig, aus Noten (als Rangwerten) arithmetische Mittelwerte zu berechnen.

durch den Einfluss der sozialen Herkunft, den sie doch nivellieren sollen (vgl. oben → Kap. 0.3).

Das Selektionssystem ist zudem ökonomisch ineffektiv, weil es Kompetenzressourcen verschenkt. Seine Kopplung an Noten führt zudem nicht zu einer trennscharfen Zuordnung von SchülerInnen zu den verschiedenen Schulformen der Sekundarstufe I. Deren Fähigkeiten überlappen sich erheblich – zumindest wenn man die Testleistung als Maßstab nimmt, wie u. a. die PISA-Daten belegen.

Damit ist die Gerechtigkeitsfrage gestellt. Denn dass Noten ihre Funktion als Selektionsinstrument nicht wirksam erfüllen, ist nur die eine Seite der Medaille. Zugleich verletzen sie auch das Recht des einzelnen Kindes auf Chancengleichheit und bestmögliche Förderung seines individuellen Potenzials. Die Kritik der ›National Coalition für die Umsetzung der UN-Kinderrechtskonvention in Deutschland‹ (2005) am schulischen Bewertungssystem macht sehr deutlich, dass eine nur systemimmanente Bewertung der Effektivität von Noten zu kurz greift:

»Die im Vordergrund internationaler Kritik stehende Bildungsbenachteiligung durch soziale Ungleichheit ist nicht nur Ausdruck eines strukturellen Mangels an Chancengerechtigkeit im gegliederten Schulsystem Deutschlands, sondern untergräbt das Recht auf Bildung jedes einzelnen betroffenen Kindes. […]

Die Leistungsbewertung durch Zensuren als Grundlage eines Berechtigungssystems ist pädagogisch fragwürdig; es verkürzt auch den Anspruch des Kindes auf Würdigung als eigenständige Persönlichkeit. Jedes Kind hat Anspruch darauf, dass seine Leistungen an seinem individuellen Vermögen, und nicht an abstrakten Regeln gemessen werden. […]

Einseitige Orientierung an Gesichtspunkten der Verwertbarkeit führt jedoch zu einer Verkürzung der Bildungsziele, die die Subjektstellung des Kindes und dessen allseitigen Bildungsanspruch unterminiert. […]

Die Vorgaben der Lehrpläne führen in Verbindung mit dem Bewertungs- und dem gekoppelten Berechtigungssystem in Deutschland zu einer weitgehenden ›Enteignung des Lernens‹ durch Fremdbestimmung.« (a. a. O., 2, 6)

Mit dem letzten Teilsatz nimmt die National Coalition ausdrücklich Bezug auf Bildungsstandards, die keine zureichende ›Offenheit‹ für die individuell unterschiedliche Entwicklung von Kindern gewährleisten und die Leistungsbeurteilung zudem auf wenige als zentral erachtete Dimensionen fokussieren. Gleiche Anforderungen für alle zum selben Zeitpunkt verletzen das »Recht auf Eigenaktivität und Selbstbestimmtheit des Kindes« (ebda).

Eine Diskussion der Noten nur als ›nützliches‹ oder ›nötiges‹ Mittel der Leistungsbeurteilung greift demnach zu kurz. Problematisch werden Zensuren durch ihre Instrumentalisierung als Auslesefilter. Der Verweis der National Coalition auf die UN-Kinderrechtskonvention macht die gesellschaftspolitische und völkerrechtliche Dimension der Notenfrage unmissverständlich klar:

»Die ausdrückliche Hervorhebung, dass das Recht des Kindes auf Bildung ›auf der Grundlage der Chancengleichheit‹ zu verwirklichen sei, unterstreicht,

dass Deutschland in diesem Punkt nicht nur bildungspolitisch, sondern auch völkerrechtlich im Abseits steht.« (a. a. O., 2)

Damit wird aber auch deutlich, dass eine ›Reparatur‹ technischer Schwächen von Noten nicht ausreicht, um die Probleme der Leistungsbewertung zu lösen. Sicher: Verbalgutachten können Leistungen, ihre Ursachen und konkrete Fördermöglichkeiten differenzierter ausweisen. Als entwicklungsorientierte Beschreibung von Lernverläufen machen sie Fortschritte und damit die individuelle Leistung des einzelnen Kindes besser sichtbar als eine Benotung im Vergleich mit anderen. Die Einbeziehung verschiedener PrüferInnen und auch standardisierter Aufgaben können helfen, die Validität, Objektivität und Reliabilität von Beurteilungen zu verbessern, indem sie informelle Leistungsproben ergänzen. Der punktuelle Einsatz normierter Tests ermöglicht LehrerInnen zudem, die vergleichende Bewertung von Leistungen auf klassenübergreifende Stichproben zu beziehen und damit ihre eigenen Maßstäbe zu überprüfen.

Eine andere Bedeutung und Wirkung gewinnen Bewertungen aber erst, wenn sich ihre Funktion ändert. Solange die Selektionsfunktion im System dominiert, werden eine stärkere Motivation der leistungsschwächeren SchülerInnen und eine differenziertere Förderung ihres Lernens nicht erreicht werden können. Auch darum ist eine längere gemeinsame Schulzeit geboten, wie sie international längst Standard ist.

Dann allerdings muss sich der Unterricht stärker öffnen für die Unterschiede zwischen den Kindern und ihnen Räume bieten für individuelle Lernwege und für eine Mitverantwortung der gemeinsamen Arbeit. Notwendig ist also eine umfassende Strategie der Schulreform. Die Individualisierung der Lernmöglichkeiten in einem sozialen Erfahrungsraum ist zu verbinden mit dialogischen Formen der Leistungsbeurteilung, die Selbst- und Fremdbewertung in den Lernprozess integrieren. Die Entwicklung geht damit über die traditionelle Frontlinie ›Verbalgutachten‹ vs. ›Ziffernnoten‹ hinaus: Standortbestimmungen, Portfolios, Lerntagebücher, Selbsteinschätzungsbögen, Lerngespräche und -vereinbarungen sind Mittel der Leistungsbewertung, durch deren Einführung sich nicht nur das äußere Format der Beurteilung, sondern auch die Beziehung zwischen Lehrenden und Lernenden grundlegend verändern kann.

Dass und wie eine solche Reform erfolgreich umgesetzt werden kann, wenn sie sich nicht auf Veränderungen der äußeren Struktur beschränkt, zeigt beispielhaft das deutschsprachige PISA-Siegerland Südtirol. Obwohl Italien insgesamt bei PISA-2003 (Lesen) mit 476 Punkten noch schlechter abgeschnitten hat als Deutschland mit durchschnittlich 491 Punkten, erreichte die autonome Provinz Südtirol bei gleicher Schulstruktur mit Platz 1 im Lesen und Platz 5 in Mathematik ein deutlich besseres Ergebnis als der deutsche Spitzenreiter Bayern. Gleichzeitig verbesserte sich die Provinz gegenüber der IEA-Lesestudie (Anfang der 1990er Jahre) von einem Platz im Mittelfeld auf einen europäischen Spitzenplatz und schneidet im Lesen noch einen Punkt besser ab als der bildungspolitische

Wallfahrtsort Finnland – mit vollständiger Integration aller behinderten Kinder, ohne Sitzenbleiben und ohne Ziffernnoten, stattdessen mit individuellen Aufgaben in offeneren Unterrichtsformen und einer Bewertung, die sich am persönlichen Lernfortschritt orientiert. Erfolgreicher Unterricht ist also auch mit weniger Leistungsdruck möglich, und Schulsysteme können lernen, ohne Selektion auszukommen.

## *Zum Abschluss: Vier Resümees aus vier Perspektiven*

Wie bei allen pädagogischen Fragen (und sozialen Phänomenen generell) ist die Befundlage zu Noten nicht auf einen einfachen Nenner zu bringen. Formen der Leistungsbewertung wirken unterschiedlich, je nachdem wie und in welchem Kontext sie eingesetzt werden. Für Folgerungen aus dem Forschungsstand kommt es deshalb darauf an, von welcher Basisannahme man ausgeht: Wer die Beweislast für Veränderungen bei den Reformern sieht, kann zu einer anderen Einschätzung kommen als jemand, der normativ die Förderung des Einzelnen als zentrale Norm und noch uneingelöste Aufgabe der Schule sieht. Vor diesem Hintergrund lässt sich als Ergebnis unserer Analysen festhalten:

- Wer an Ziffernnoten festhalten will, weil sie angeblich objektiv und vergleichbar seien bzw. erforderlich, damit SchülerInnen sich auf die Anstrengungen des Lernens einlassen, findet in der Empirie keine stützenden Belege für seine Position.
- Auch diejenigen, die Verbalgutachten ablehnen, weil sie angeblich negative Auswirkungen auf die Lernbereitschaft und den fachlichen Lernerfolg der SchülerInnen haben, können sich auf keine empirischen Daten stützen.
- Wer andererseits hofft, ohne zusätzliche Maßnahmen, d.h. allein durch die Verordnung von Verbalgutachten Lernbereitschaft und Lernerfolg von SchülerInnen verbessern zu können, wird durch die Befunde zur bisherigen Beurteilungspraxis und ihre Wirkungen ernüchtert. Ohne eine pädagogische und didaktische Öffnung des Unterrichts und ohne die Sicherung bestimmter Rahmenbedingungen bleibt eine Veränderung der Bewertung meist erfolglos.
- Diejenigen aber, die mit dem Verzicht auf Ziffernnoten pädagogische Ziele verfolgen, können mit einer Verbesserung der Unterrichtssituation und des Lernerfolgs, vor allem der schwächeren SchülerInnen, rechnen – sofern sie bereit sind, als LehrerInnen einen höheren Aufwand zu leisten, als Schulverwaltung mehr in die Fortbildung und Unterstützung der LehrerInnen zu investieren und als BildungspolitikerInnen den Selektionsdruck im System zu verringern.

## 8. Ausgewählte Literatur zur Vertiefung

Für die praktische Umsetzung besonders hilfreiche Publikationen sind mit * gekennzeichnet


* Bambach, H., u. a. (Hrsg.) (1996): Prüfen und beurteilen. Zwischen Fördern und Zensieren. Jahresheft XIV. Friedrich-Verlag: Seelze.

* Bartnitzky, H. u. a. (Hrsg.) (2005): Pädagogische Leistungskultur: Materialien für Klasse 1/2. Beiträge zur Reform der Grundschule, Bd. 119. Grundschulverband: Frankfurt.

* Becker, G. u. a. (2006): Diagnostizieren und Fördern. Stärken entdecken – Können entwickeln. Jahresheft XXIV. Friedrich Verlag: Seelze.

Beutel, S.-I. (2004): Zeugnisse aus Kindersicht. Habilitation an der Universität: Jena (publ. 2005 in der Schriftenreihe der Max-Traeger-Stiftung. Juventa: Weinheim / München).

Beutel, S.-I. / Vollstädt, W. (Hrsg.) (2000): Leistung ermitteln und bewerten. Bergmann + Helbig: Hamburg.

Brügelmann, H. (2005): Schule verstehen und gestalten – Perspektiven der Forschung auf Probleme von Erziehung und Unterricht. Libelle: CH-Lengwil, Kap. 56 ff. Fortlaufend aktualisiert unter → www.agprim.uni-siegen.de/schuleverstehen

Ingenkamp, K. (Hrsg.) (1995): Die Fragwürdigkeit der Zensurengebung. Beltz: Weinheim (1. Aufl. 1971).

Jachmann, M. (2003): Noten oder Berichte? Die schulische Beurteilungspraxis aus der Sicht von Schülern, Lehrern und Eltern. Leske + Budrich: Opladen.

Lübke, S.-I. (1996): Schule ohne Noten. Lernberichte in der Praxis der Laborschule. Leske + Budrich: Opladen.

Maier, M. (2001): Das Verbalzeugnis in der Grundschule. Verlag Empirische Pädagogik: Landau.

Valtin, R. (Hrsg.) (2002): Was ist ein gutes Zeugnis? Noten und verbale Beurteilung auf dem Prüfstand. Juventa: Weinheim/München.

* Winter, F. (2004): Leistungsbewertung. Eine neue Lernkultur braucht einen anderen Umgang mit den Schülerleistungen. Schneider Hohengehren: Baltmannsweiler